コーパスから見える文法 *

大名　力 **

Corpora and Grammar

Tsutomu OHNA

Abstract

An empirical investigation of a grammar, a part of the internal state of a speaker, should be conducted based on externally observable data. With the sizes of corpora increasingly large and the expansion in the amount and kind of data obtainable from them, corpora are now more and more widely used for hypothesis testing. Such corpus data sometimes reveal new facts overlooked so far, leading to the falsification of a widely accepted hypothesis in favor of another which may be counterintuitive.

In spite of their usefulness, corpora can be easily abused: With the development and spread of "user-friendly" environments, users tend to pay attention only to the output of software while disregarding the input and the process and not examining whether the data, especially statistical ones, can be interpreted appropriately as evidence for their hypotheses. For the development of corpus-based research, we should examine not only the validity but also the limitations of present methods and potential problems which may be posed by them, as well as often hidden assumptions concerning use of corpora in linguistic research.

0. はじめに

母語話者の内部状態の一面としての文法は直接観察できないため，文法を用いた活動の結果生じる観察可能なデータを基に仮説検証を繰り返すことで探っていくしかない．長い間，書籍や新聞等の印刷物，テレビ・ラジオなどの放送番組などから収集された言語資料，実験や内省から得られるデータなどを中心に研究が進められてきたが，最近ではコーパスの利用も進んでいる．活動の種類により反映されやすい文法の側面も異なるが，コーパスに限ってみても，種類や規模により得られる情報が大きく異なる．本稿では，大規模コーパスに焦点を当て，大規模コーパスの利用により文法のどのような側面がよりよく観察できるようになったのか，また，それは仮説検証にどう役立つのかについて見た後，コーパスを有効活用するために注意すべき点について考察する．

* 本稿は，日本言語学会第 136 回大会招待講演「コーパスから見える文法」予稿集原稿（『日本言語学会第 136 回大会予稿集』, pp. 36–45）に大幅な加筆修正を施したものである．

** 国際開発研究科国際コミュニケーション専攻准教授．

1. コーパスと文法の関係

例えば，*wildly* という副詞がどのような形容詞を修飾しやすいか調べたいとする．このような場合，従来であれば，辞書・語法書を見る，母語話者に聞く，本・新聞などから用例を収集するなどの方法を採っていたが，現在では，コーパスを利用するケースも増えている．（Cf. 滝沢 2009）(1) は，筆者がウェブサイトから収集した The Voice of America の放送原稿（約 1 千万語）を対象に *wildly* を検索し，右の文脈でソートし出力した結果である．(2) は，The Bank of English というコーパス（5 億語以上）で，“wildly ＋形容詞”を検索し，*wildly* の直後に来る語（頻度順上位 10 語）の統計値（頻度，t-score，MI-score）を示したものである[1)]．どちらのデータからも，*popular*，*successful*，*different* などの形容詞が *wildly* により修飾されやすいことがわかる．このように，コーパスを利用すれば，手作業では時間がかかり困難か，あるいは不可能な処理が比較的容易に行える．

(1)

```
1.   ical-scan ballots varied wildly -- and not just from county t
2.   argely unpainted face is wildly accented with exaggerated lip
9.   or prevents death from a wildly beating heart. VOICE TWO: Doc
10.  // That turned out to be wildly bullish - in November of 1994
11.  lim cleric Mr. Wahid was wildly cheered as he took over the p
12.  independent media issued wildly contradictory claims about th
13.  rate for the currency is wildly different than the government
14.  NNEY :16 "The process is wildly different. You have very litt
15.  Thirty Days in Sydney: A Wildly Distorted Account," and it's
16.   those refund claims are wildly exaggerated. The chairman of
17.  t. The Corvette Z-O-6 is wildly impractical, not cheap at nea
18.  ts of Washington, D-C is wildly inappropriate. There's no pla
24.  ngs of vulnerability are wildly overblown, the reality that y
25.  r. 1 // REST OPT /// The wildly popular former guerilla leade
26.  r of Antananarivo and is wildly popular in the capital city.
27.   continues. The mayor is wildly popular in the capital, and m
28.   Ladies' figure skating, wildly popular in the United States,
29.  -D-P is hopeful that the wildly popular prime minister will h
30.  tmas classic the new and wildly popular theatrical movie "Dr.
31.  e industry which remains wildly successful despite Japan's ec
32.  HARBOR. EVERYONE CHEERED WILDLY WHEN A SIGNAL WAS GIVEN AND A
```

(2) Query: “wildly ＋ ADJ”, 1647 matching lines

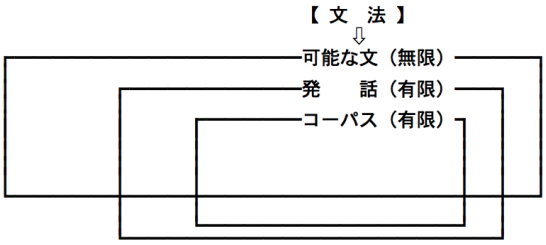
	各語の頻度	「widly ＋形容詞」の頻度	t-score	MI-score
popular	56573	120	10.9382	9.4006
successful	52461	90	9.4695	9.0944
different	167391	71	8.3638	7.0782
optimistic	9589	63	7.9335	11.0319
inaccurate	2215	56	7.4824	12.9762
enthusiastic	7560	52	7.2078	11.0980
fluctuating	760	34	5.8305	13.7996
expensive	30109	33	5.7281	8.4480
improbable	1967	32	5.6558	12.3401
wrong	80872	23	4.7429	6.5015

あるコーパスでヒットしなくても，一般に使われない表現だと断言することはできない．先の The Voice of America の放送原稿では，*wild popularity*，*wild success*，*wild difference* は出現しないが，The Bank of English では，*wild difference* を除き，該当例がヒットする．問題の表現がヒットするかどうかは，コーパスのサイズ等によっても変わってくるため，使用するコーパスの内容をよく把握しておく必要がある．「コーパス」と言ってもいろいろなものがあるが，本稿では「大規模な，電子的に処理可能な，主として（標準的な変種の）成人母語話者の，文字で書かれた（書き起こされた）自然発話の集積」としての「コーパス」を扱うことにする．具体的には，The Bank of English と The British National Corpus を使用する．どちらも，書き言葉が主で，品詞情報が付与されている．また，適宜，ウェブページから採取した例も利用する[2)]．

コーパスの規模が変われば得られるデータの量・種類も変わる．世界初の電子コーパス，通称 Brown Corpus の完成（1964 年）から半世紀近くが過ぎているが，コーパスが広く言語研究に利用されるようになったのは比較的最近のことである．コーパスの利用が一般的でなかった理由の 1 つは，コンピュータそのものが普及していなかったことにあるが，それに加え，60 年代〜 80 年代に利用可能であったコーパスは，現在からすれば規模が小さく，得られるデータの規模・範囲も限られていたため，仮説の検証に必要なデータが得にくかったことも理由の 1 つであろう．しかし，コーパスの大規模化，処理ツールの高速化・高機能化により，コーパスを利用することで，これまで見過ごされてきた新しい事実の発掘や，それらの事実に基づく仮説の検証が可能な状況になってきている．

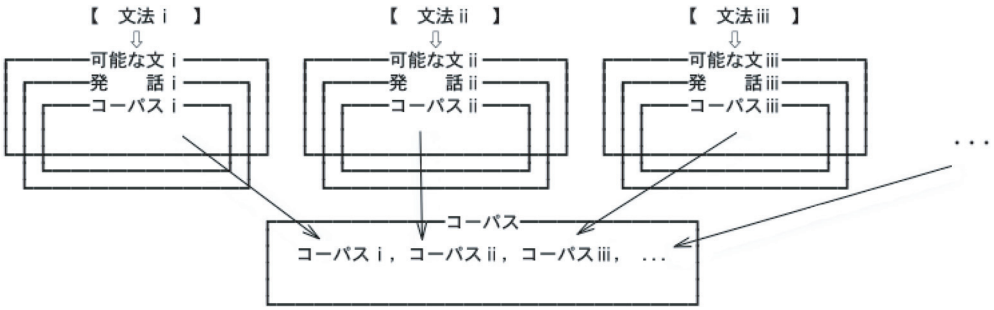
文法，文法的な文の集合[3)]，発話，コーパスに収録される発話の集合の関係は（3）のようになる．実際の発話には，言い誤り，不完全な文等，文法的でない文も含まれるため，発話を収集したコーパスにも，文法的でない発話が含まれる可能性がある．

(3)



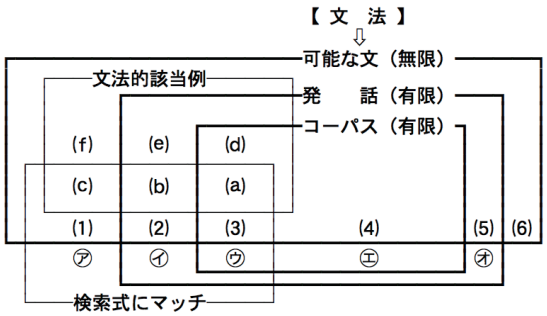
(3) は個人差を捨象した図となっている．大規模コーパスでは複数の話者が産出した文を収録し，個々人の発話のうちコーパスに収録されたもの全体を「コーパス」と呼ぶため，個人の脳内に存在する文法との関係は厳密には (4) のようになるが[4)]，ここでは便宜的に (3) のように簡略化して示す．

(4)



可能な文，実際に発話された文でも，コーパスに収録されなければ利用できないのは当然だが，収録されていても取り出せなければ利用はできない．手作業では不可能な量のデータを処理できるのが大規模コーパス利用の利点の１つであり，文の抽出はコンピュータによる処理に頼らざるをえないため，コーパスへの付加情報，使用するツール，検索方法が重要になる．「可能な文」の集合のうち，当該の言語現象に関わる文の集合を「文法的該当例」，検索式にマッチし取り出せる部分を「検索式にマッチ」として示すと，次のようになる[5]．

(5)



主として研究に利用できるのは（a）の部分であり，文法研究に役立つためには，（a）が質・量ともに十分でなければならない．比較的最近までは十分とは言えない状況であったが，今は状況も大きく変わっている．以下，具体的に，コーパスの大規模化により，現在どのようなデータが利用可能になっているのか，また，それが，仮説検証にどのように役立つのかについて見ていくことにする．

2. コーパスデータによる仮説の検証

本節では，例として，*a beautiful two weeks* のような名詞句（以下，便宜的に「ABTW 構文」と呼ぶ）と，*day after day* のような名詞句（「N *after* N 構文」）を取り上げて見ていく．

2.1 ABTW 構文

an estimated fifty students のような「不定冠詞＋形容詞（相当）句＋数詞＋複数名詞（可算名詞）」

からなる名詞句は，複数名詞に不定冠詞が付いている特異な表現として，伝統文法の頃から研究者の注意を引いてきた．辞書や文法書では，*full*, *further, estimated* などを扱った項目で触れられていることが多いが，(6) に示したように，一部の形容詞の例外的な用法というわけではなく，現代英語では極めて生産性の高い構文である．

(6) a *thin* 300 million shares, a *scant* two inches, a *record high* 3.2 percent, a *miserable* thirty-three dollars, a *brief* six minutes, a *whopping* 540 pages, a *possible* 1,600 yards, an *incredible* 575 tokens, a *staggering* 29 goals, an *inherently more amazing* five Wimbledon titles, an *exceptional* 125 rebounds, a *crucial* seven points, a *very ambitious* 60 minutes, a *murderous* thirty minutes, a *delightful* two hours, an *uncomfortable* 10 days, a *humid* 80 degrees [BoE]

複数名詞と不定冠詞の共起は，Jespersen（*MEG* II. §§ 5. 11–18）の「複数の統合」（unification of plurals）またはそれに類した考えによって説明されることが多い．「複数の統合」とは，複数の物のまとまりを 1 つの，より高次の単位と捉え，単数として扱うことを言う（cf. 荒木•安井（編）1992：1533）．Jespersen 自身は明言していないが，名詞句全体が 1 つのものとして扱われ不定冠詞が付くと言っているので，概略 (7a) のような構造をしていると考えてよい．これが一般的に受け入れられている分析（以下，「複数の統合」説と呼ぶ）だが，Jackendoff（1977：128–130）は，これとは異なる分析を採り，概略 (7b) の構造，すなわち，不定冠詞が形容詞とともに数詞を修飾する構造を提案している．（「数詞の修飾句」説と呼ぶ．）

(7) a. 「複数の統合」説： [NP a [beautiful two weeks]]
b. 「数詞の修飾句」説： [NP [a beautiful two] weeks] cf. [NP [a dozen] weeks]

「数詞の修飾句」説では，不定冠詞の生起は名詞句全体の数の解釈とは無関係ということになり，事実に関し「複数の統合」説とは異なる予測をするため，事実と照らし合わせることで，どちらが妥当か検証することができる．以下，簡単に「数詞の修飾句」説を支持する証拠を見ることにする．（詳細は Ohna（2003）を参照．）

「複数の統合」説によれば，問題の名詞句は単数名詞として振る舞うことが期待される．動詞との数の一致，代名詞の数に関する事実を見ると，確かに，単数扱いされることはあるが，複数扱いされることも多く，不定冠詞が付いているからと言って単数としての読みが強制されるわけではないことがわかる．

(8) a. An estimated 55 pence is spent on packaging. [BNC]
b. He's announced that an extra seventeen million pounds is to be pumped into care over the next six years. [BNC]
c. In 1969 alone, an estimated 166,000 dolphins were killed by Turkish hunters, most of them shot. [BNC]
d. A further 30,100 people were out of work because they had been stood down, up from 26,600

the previous month. [BoE]

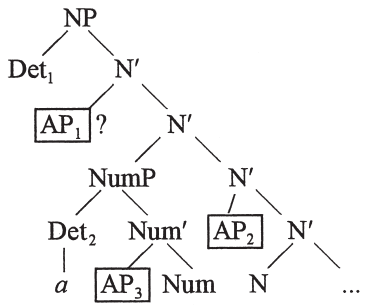
また，次のように，意味的に複数名詞が要求される（そう解釈するのが自然な）文脈に生起可能なことから，（主要部の名詞が単数である場合を除き）問題の名詞句が意味的に単数として扱われていると考えることには無理がある．

(9) a. Of an arbitrary 13 goods, five were more expensive in the independents, seven more expensive in Sainsbury's, and one (tinned tomatoes) cost the same. [BNC]
b. ... so that now, the tennis shoe will consist of an estimated 12 separate components. [BNC]
c. In a single, one-hectare plot in Peru, one botanist counted 600 trees, divided into an astonishing 300 species. [BoE]
d. The Environmental Protection Agency has identified 123 U.S. plants - including several near Philadelphia - that, in a serious explosion, could harm at least a million nearby residents. An additional 700 plants could each endanger 100,000 people. [BoE]
e. A further 50 students began to arrive at the gate of the hall.

これらの事実は「複数の統合」説では不定冠詞の出現を説明できないことを示している．「数詞の修飾句」説では，不定冠詞は名詞句全体の数の解釈とは無関係であるため，これらの事実とも矛盾しない．

「数詞の修飾句」説を積極的に支持する統語的な証拠もある．この分析によれば，名詞句の決定詞の位置（(10) の Det_1 の位置）は空いていることになり，そこに別の冠詞が生起することを予測する．

(10)



(11) a. [[*(a) dozen] books]
b. [the [(*a) dozen] books]
c. [[(a) few] books]
d. [the [(*a) few] books]

(12) a. [[*(an) estimated 80] people]
b. [the [(*an) estimated 80] people]

英語では冠詞の連続は認められないため（cf. (11)），(12b) でも不定冠詞があると容認不可となるが，次の例のように，冠詞が連続しなければ共起は可能である．

(13) a. The 2 to a possible 4 positions offered by the center are only part-time jobs.
b. [NP the [2 to a possible 4] positions ...]

(14) a. ... compared to the more than an estimated $7 billion spent by Medicare in FY 1997 for graduate medical education to support physician residencies in hospitals, ... [Web]

b. Arrow and Salt are just two of the more than an estimated 1500 humpback whales in this area. [Web]
c. ... the up to an extra two per cent of the declared distribution they received in 2006 is generally considered to be income for tax purposes. [Web]

(15) a. [NP the [more than an estimated 7 billion] dollars ...]
b. [NP the [more than an estimated 1500] humpback whales ...]
c. [NP the [up to an extra two] per cent ...]

「数詞の修飾句」説であれば，冠詞の共起も，また，全体が定名詞句であるに拘わらず不定冠詞が生起することも問題にはならない．

このように，コーパスから，一見直感に反する「数詞の修飾句」説を支持するに足る十分なデータが得られる．

2.2 N *after* N 構文

次に N *after* N 構文について見てみよう．この構文に関する研究はいくつかあるが，ここでは，関連構文も含め，この構文の統語的・意味的特徴を詳細に記述している Jackendoff（2008）を取り上げ，N の同一制約に関わる部分に焦点を当て見ていく．（詳しくは Jackendoff（2008）とそれに挙げられている文献を参照のこと．）

2.2.1 Jackendoff（2008）の分析

まずは基本的な事実を確認することにする．次の（16a，b）の対比から，N *after* N の 2 つの N は同一であること，また，(16c)から，N が 2 つ現れるパターンだけでなく，“N *after* N *after* N”と，さらに“*after* N”を追加していくことが可能なことがわかる．

(16) a. I read book after book.
b. *I read book after newspaper.
c. I read book after book after book.

前位修飾の形容詞については，(17a，b) のように，最後の名詞のみか，すべての名詞に付くのが普通であり，(17c，d) のように，最初の名詞以外に付いたり，最初の名詞のみに付くのは非文法的とされる．但し，(17e) のように，このパターンを破っても適格とされるものもある．(17e) の解釈に関し，Jackendoff は (18) のように述べている．

(17) a. week after week after miserable week
b. miserable week after miserable week after miserable week
c. *week after miserable week after miserable week
d. *miserable week after week after week
e. week after miserable week after thoroughly rotten week (Jackendoff 2008: 21)

(18) Curiously, the sense of 40e [= (17e)] is that the weeks get successively worse, while the literal meaning of *after* suggests that the worst week comes first. (*ibid*.)

これらの事実を基に，Jackendoff は N *after* N 構文を，(19) に示したような，意味―統語―音韻の複合体として分析している．

(19) Meaning MANY (MOD_k) X_is IN SUCCESSION
a. Syntax $[_{NP}\ N_i\ P_j\ (A_k)\ N_i]$
Phonology Wd_i $after_j$ (Wd_k) Wd_i
b. Syntax $[_{NP}\ (A_k)\ N_i\ P_j\ (A_k)\ N_i]$
Phonology (Wd_k) Wd_i $after_j$ (Wd_k) Wd_i (*ibid*.)

(19a) は最後の名詞にのみ，(19b) は全ての名詞に形容詞が付くケースに対するものである．Syntax における 2 つの N に付いている指標 i により，2 つの名詞が同一であること，Meaning の X_i，Phonology の Wd_i に対応することが指定される．b. の形容詞についても同様である．ところが，前置詞の *after* に関しては，Meaning には直接対応するものがない．これは，(17e) のように，*after* 本来の意味が薄れていると思われる例を考慮してのことである．さらに Jackendoff は，英語では Syntax に最初の N を設ける必要性はないとし，(20) のように分析する可能性も示している．

(20) Meaning MANY X_is IN SUCCESSION
Syntax $[_{NP}\ P_j\ N_i]$
Phonology Wd_i $after_j$ Wd_i (*ibid*.: 26)

以上，Jackendoff の分析を簡単に見たが，彼の分析の特徴のうち，(21) に挙げた点には問題があり検討が必要であるが，ここでは，(21c) の，名詞が同一でなければならないという制約（以下，「同一名詞制約」と呼ぶ）およびその制約の指定の仕方を取り上げて考えてみたい．

(21) a. N *after* N 構文の変種を区別せずに扱う
b. 意味には，音韻構造の $after_j$，統語構造の P_j に直接対応する物は現れない
c. i. 2 つの N，A はそれぞれ指標により同一であることが指定されている (← (19))
ii. 最後の N，A 以外は統語構造には現れず，音韻構造のみに現れる (← (20))

2.2.2 “同一名詞制約”と形容詞による修飾

Jackendoff は，N *after* N 構文に現れる複数の N と A は，それぞれ同一であるとしているが，Jackendoff 自身が挙げている *week after miserable week after thoroughly rotten week* がこの条件に合わない．また，名詞についても，コーパスやサーチエンジンを利用して調べてみると，異なる名詞が *after* の前後に現れる例が見つかる．

(22) a. But today girl after boy, boy after girl, says “not prepared”; ... [*The English Journal*, Sept., 1940, NCTE]

b. Wave upon wave of children flocked, boy after girl, girl after boy, but only the young wealthy ones were permitted. [Web]
c. Student after teacher after student approached the podium with kind words ... [Web]

(23) a. ... with study after story after column about how dumb, greedy, and just plain bad they supposedly are. [BoE]
b. Then she was away, making walk after border after summer house, ... [BoE]
c. Well-intentioned parent after student after teacher after interested member of the public ... expressed themselves, ... [Web]
d. How big is the Burcham family? I met auntie after cousin after brother after sister! [Web]

名詞が同一でなければならないとすると制約がきつすぎ，逆に，同一でなくともよいとすると，*book after newspaper* のような例の不適格性が説明できなくなるという問題が生じ，同一名詞制約が適用されるケースとされないケースを区別して扱う必要がある．

N *after* N 構文における名詞の異同と文法性の対応を整理すると，(24) のようになる．(24) のパターンからは，N が同一でなければならないのは，N が 2 つの時ということがわかる．

(24) a. i. A → A　　student after student
ii. A → A → A　　student after student after student
iii. A → B, B → A　　student after teacher, teacher after student
b. A → B → A　　student after teacher after student
c. i. *A → B　　*student after teacher
ii. A → A → B　　student after student after teacher
iii. A → B → C　　student after teacher after parent

N が 2 つと 3 つ以上の時とを分け，2 つの時にのみ同一指標を付ければ事実には合うが，2 つと 3 つ以上で制約が異なることは，単なる偶然として扱われることになってしまう．

では，(24) の c.i. と他を区別する要因は何か．それは，人間が「パターン」を読み込めるかどうかであると思われる．直感による判断であるが，通常，最低 3 つのもの（変化・推移としては 2 回）が連続しないと，パターンを成すとは見なさないように思われる．

(25) a. A → A　　A↻A　　A → A → A（→ A → A ...）
b. A → B, B → A　　A⇄B　　A → B → A（→ B → A ...）
c. A → B → A　　A↻B, A↺　　A → B → A（→ A → B → A ...）
d. A → A → B　　A↻A, B↺　　A → A → B（→ A → A → B ...）
e. A → B → C　　i. A↻B, C↺　　A → B → C（→ A → B → C ...）
ii. A↻B↻C↷ ...　　A → B → C（→ D → E ...）

このような観点から考え直してみると，“N *after* N” と N が 2 つのみの場合，3 つ以上の連続と

して展開されるには，N が同一で再帰的に解釈される場合のみということになり，N の数が 2 つと 3 つ以上で異なることも，単なる偶然ではなくなる[6]．

元々ある“同一名詞制約”が，ある条件下で外れるのか，それとも，最初からそのような統語・音韻的な制約はないのかについては，さらに検討が必要であるが，N *after* N 構文の適格性，解釈に人間のパターン認識能力が関わっていることは間違いないであろう．

2.3 仮説検証における内省とコーパスによるデータ

Jackendoff が「数詞の修飾句」説を提案してから 30 年以上が過ぎているが，この説が話題となることはほとんどない．これはなぜであろうか．ABTW 構文は比較的出現頻度が高く，珍しい表現であるため話題とならないとか，また，Jackendoff（1977）を読む人が稀で説自体が人の目に触れていないというのは，理由として考えにくい．おそらく，直感に反した分析であることが理由の 1 つであったと思われる．直感に反していても，仮説検証に十分なデータが利用できれば状況は違っていたかもしれないが，1970～80 年代に利用可能であった 100 万語レベルのコーパスでは，出現するバリエーションの種類が少なく，仮説の検証に必要なデータを集めることは難しく，また，新聞・書籍等からの手作業による用例収集では，時間がかかるだけでなく，出現する例も単純な例が多いため，仮説の検証へと繋がらず，仮説自体が省みられなくなってしまったのではないか．

内省の利用もデータ収集の有効な手段の 1 つだが，実例を見て初めて意識化されることも少なくない．Jackendoff（1977：130）は，*a beautiful one day* を非文法的とし，*one* を他の数詞と別扱いしているが，実際には *a mere one game* のように *one* もこの構文で用いられるため，*a beautiful one day* の不適格性は，形容詞による修飾可能性とは別の要因によるものと考えた方がよい．N *after* N 構文についても，大規模コーパスで検索して確認してみれば，N が 3 つ以上現れるタイプであれば，同一名詞でない例は比較的容易に見つかり，同一名詞制約はすべての変種に当てはまることではないことがわかる．すぐれた言語感覚を持つ母語話者であっても，このような単純な事実に気付かないこともある．作例を利用する利点の 1 つに，最小対を比較することにより要因を特定しやすいということがあるが，単純な最小対の比較から誤った帰結を導き出してしまうこともある．

(26) a. a beautiful two days
b. (*)a beautiful one day
c. a mere one game

(27) a. study after study
b. (*)study after story
c. study after story after column

手作業では手間のかかることが，コーパスを利用すれば簡単に済むことも多い．また，コーパスの検索結果が，時に予期せぬデータを提示し，新たな視点を与えてくれることもあり，内省を利用する場合でも，コーパスを援用することは望ましいことである．とりわけ，母語でない言語を研究する者にとって，短期間に信頼性の高いデータを多量に収集することが可能になるため，メリットは非常に大きい．

コーパスは，出現頻度の低い周辺的な構文の研究にのみ役立つものであると考える人もいるかもしれないが[7]，コーパス利用の有効性は周辺的な構文の研究のみに限定されるものではない．基本形の頻度は高いが，周辺的な変種の頻度は低いという場合もある．基本形に関する事実はどの理論でも扱えるようになっているのが普通で，周辺的な変種としてどのようなものが可能なのか，また，それらがどのような属性を持つかが，仮説の妥当性を決める時に重要になることも少なくなく，コーパスから得られる多様な変種に関するデータは，基本的な構文・構文の基本形を分析するうえで重要な役割を果たすことになり，構文の分析の見直しにより，他の構文，文法全体の再検討が必要となることもある．一般に解決済みと考えられているものであっても，コーパスを利用し再度検討してみる価値は十分にあろう．

3.「コーパスから見える文法」

この節では，1・2 節で見たことを踏まえ，次の 3 つの観点から「コーパスから見える文法」について考えてみたい．

(28) A. コーパスデータから見える文法の実態
B. コーパスデータから見える文法の側面，見えない側面
C. コーパスデータを使って初めて見えてくる文法

3.1 A. コーパスデータから見える文法の実態

派生的な変種や文法操作が加わった例などは頻度が低く，小規模なコーパスでは出現が期待できないことが多かったが，コーパスの大規模化，品詞などの情報付加，処理ツールの高速化・高機能化により，そのような例もコーパスから得られるようになり，コーパスデータによる仮説検証も容易になってきた．例えば，ABTW 構文で，(29) のような，冠詞の共起の確認には，従来であれば，実質的に作例に頼る以外に方法はなかったであろう．

(29) a. the [more than an estimated 7 billion] dollars
b. the [more than an estimated 1500] humpback whales
c. the [up to an extra two] per cent

また，構文の基本形だけでなく派生的な変種も抽出できるようになったことにより，基本形から変種，変種からより派生的な変種への拡張のパターンも捉えやすくなってきている．N *after* N 構文で，*student after teacher after parent* のように異なる名詞が出現するケースは，5 億を超える語数の The Bank of English でも 4 例程度しかヒットせず，小規模コーパスに類例が含まれている可能性は極めて低いが，ウェブページも含め，大規模なテキストを対象とすれば，かなりの数の例が集まり，これらの変種も含めて，構文全体の再検討が可能になる．

構文の拡張のパターンを検討していると，認知的，機能的要因が関わっていると思われるケー

スが出てくる．上で見たように，N *after* N 構文では，N の数が 3 以上の時には同一名詞という制約が外れる．単なる事実の記述であれば，統語的（＋局所化可能な意味的）な道具立てだけでも処理することはできる．しかし，なぜそのような変種が存在可能なのかに答えようとすれば，人間のカテゴリー形成，パターン認識のような認知能力を考慮しなければならなくなる．このように，コーパスを利用し多様な言語事実を見ることにより，基本的な事実のみからは見えにくかった文法の実態が見えてくるようになる．

3.2 B. コーパスデータから見える文法の側面，見えない側面

以前に比べ有用度が高くなったと言っても，コーパスは万能ではなく，コーパスデータによって調べやすい文法の側面もあれば，調べにくい側面もある．

基本的に，コーパスから否定証拠は得られない．したがって，主として肯定証拠を基に仮説の検証が行えるものの方が，コーパスから必要なデータを集めやすい．「数詞の修飾句」説が（特に，「複数の統合」説との対比で）予測する ABTW 構文に関する事実は，肯定証拠で検証できるものが多く，且つ，ABTW 構文の場合，バリエーションも豊富に出現するため，コーパスデータによる仮説の検証が容易である．（Cf. 大名 2004）

自然な発話には生じない（あるいは，生じてもコーパスに収録されにくい）文を分析に必要とする研究にコーパスが役立たないのは当然であるが，コーパス内にあっても，取り出せなければ役には立たない．現実には，検索しやすいものとしにくいものがあり，それにより調査可能な文法の側面も変わるが，一般に (30) の条件を満たすものは検索しやすい．形態素や単語を直接指定して検索できるものの方がデータが得やすいが，語法やコロケーションの研究でコーパスがよく利用されるのも，このような事情によるところが大きい．

(30) a. 特定の形態素を指定できる．
b. 直接指定可能な要素が連続する．
c. 要素間の配列順序が不変である．
d. 要素の欠落（省略），代名詞等の代用表現による置き換えがない．

英語では品詞情報も比較的利用しやすい．これは，i) 一般的に受け入れられている品詞分類がある，ii) 正書法が確立しており，「語」の認定が比較的容易である，iii) 品詞の付与はある程度自動的に行え，精度も実用的なレベルにある，などの理由に因る．

これに対し，統語情報，意味情報等の情報は利用しにくいのが現状である．（Cf. *wildly* が修飾する形容詞の「意味」，ABTW 構文における意味的修飾関係，同一指示，など．）また，「省略」や局所化できない意味を含む例の検索も困難である．N *after* N 構文で問題となったような，「パターン認識」のような概念は局所化が難しいこと，他でも広く利用される可能性の高い一般的な意味属性ではないことから，意味情報が付加されたコーパスでも，この種の情報が記載されることは期待できない．N *after* N 構文の場合には，*after* が 2 回以上出現することを利用し，問題となる例を抜き出すことができたが，そのような処理ができないものについては，検索は難しく，分

析にそのようなデータを必要とする言語現象は，コーパスに該当例が多数含まれていたとしても，検証しにくいことになる．

3.3 C. コーパスデータを使って初めて見えてくる文法

コーパスを使わないと見えてこない文法の側面があれば，コーパスの価値はより高くなるが，現時点では，コーパスならではというものがあるとは言えない状況かもしれない．

1節で見たMI-scoreを例にとって考えてみよう．MI-scoreは2語の連想関係の強さを計る尺度で，よくコロケーションの抽出に利用されるが，実は，コロケーション性を判断するはっきりとした基準値があるわけではない．例えば，齊藤他（2005：133）は次のように述べている：

(31) コロケーション性を判断する基準値は厳密に決まっているわけではない．MI-scoreについては，Church & Hanks（1990：24）が，概して3より大きい場合に面白い組み合わせになっていると述べている一方，1.58以上を目安にしている記述もある（Barnbrook 1996：99）．本章では3より大きい場合に注目することにする．t-scoreについては2以上が基準とされることが多く（Barnbrook 1996：98, Hunston 2002：72他），本章もこの値を採用する．

MI-score自体が言語知識の何かに直接対応しているわけではなく，利用されている基準値というのは，コロケーションの専門家が，経験的に，その値以上であればコロケーションとして扱いたいものが取り出せると判断した便宜的な数値である．つまり，MI-scoreの有効性は，専門家の経験により判断されているわけである．

他の統計値•手法についても，その有効性は，まずは，その処理結果が，信頼できる専門家の「感覚」と一致するかにより判断されるのが普通である[8]．その意味で，「優れた研究者でも内省により得られない情報を引き出す手法」とはならないのは当然のことである．しかし，このことによって，そのような手法の開発に意味がないということになるわけではない．信頼できる手法の開発は，客観性，「知識」の共有化，自動抽出などの点でメリットがある．信頼度が高まれば，大規模なデータへ適用した結果が，専門家でも判断できないものであっても，信頼できる結果として扱うことができるようになる．

開発の段階で，「コーパスやその手法を使わなくてもわかる結果が出るだけだから，コーパスの利用やそのような手法の開発には意味がない」と考えてしまうと，研究の価値を読み誤ってしまう可能性があるので，注意が必要である．

4. コーパスの有効活用のために

コーパスは役に立つツールであるが，その特性をよく理解して利用する必要がある．本節では，コーパスを利用する際の注意点について，いくつか例を挙げて見ていくことにする．

4.1 コーパスの内容（ジャンル，規模，内部構造，付加情報など）を把握する

紙媒体に基づく手作業の処理では，資料を見ずに作業は行えないため，必然的に対象の確認は行われるが，コンピュータによる検索では，検索対象のコーパスの内容を確認せずに検索を行ってしまう危険性が高くなる．

一口にコーパスと言っても，内容はいろいろである．話し言葉に特徴的な現象を調べるのに，書き言葉のコーパスを利用するのは不適切である．品詞情報が付加されていないコーパスでは，品詞による検索はできないし，品詞情報が利用できる場合でも，精度の問題のため，品詞を指定しない方がよいこともある．ヒット数が 0 でも，元々，ジャンルや規模の所為で，求める情報がコーパス内に存在しない可能性もある．コーパスを適切に利用するには，コーパスの内容をよく把握しておく必要がある．

4.2 入力・処理・出力をセットで考える

コーパス中に存在するデータも取り出せなければ利用はできないため，検索ツールや検索技術が重要である．コーパスへの付加情報により取り出せるデータの量・質も変わってくるため，利用にはそれらに関する知識も必要である．

最近では“ユーザーフレンドリー”なツールも増え，特別な訓練を経なくても大規模コーパスが簡単に利用できる状況となっているため，検索ツールの技術的な理解や特別な検索技術の習得は不要と思われがちであるが，特別な訓練なしで簡単に使えるツールを利用することには問題もある．通常の社会生活でコンピュータを利用する場合，目的が達成されたかどうかは結果のみから判断できるのが普通であるが，研究においては，入力・処理・出力（処理対象・処理内容・処理結果）の 3 点をセットとして考えなければ，結果が正しいかどうか判断することはできないことが多い．“ユーザーフレンドリー”なツールでは，入力と処理の部分が隠されてしまうことが多く，出力の正しさの検証が難しくなるだけでなく，そもそも，検証の必要性自体が意識されにくくなる．研究においてコーパスを適切に利用するには，面倒ではあっても，基礎的な情報処理技術の習得が不可欠である．

4.3 排除されるものが何かを意識する

コンピュータを使えば，1 億語のコーパスからも短時間で情報を引き出すことができるが，1 億語すべてを自分の目で直接観察しているわけではない．直接データを見てチェックしていれば，該当例（として考慮すべきかもしれないもの）であることに気付き，また（しばしば暗黙のうちに立てられている）仮定を修正することが容易な場合でも，コンピュータによる検索では，そのような例の存在に思い至らない可能性が高い．例えば，*a beautiful two weeks* タイプの名詞句を検索するのに，名詞として複数名詞を指定したり，不定冠詞・形容詞・数詞の連続を指定すれば，*a mere one game* のような単数名詞の例や，(32) のような例はヒットしないため，そのような例の存在に気付かない可能性が高い．

(32) a. ... “as good a 45 minutes as a side of mine has ever produced”. [BoE]
b. ... with the first half being as entertaining a 45 minutes as any this season. [BoE]
cf. a [very entertaining] 45 minutes, *an [as entertaining] 45 minutes

検索式にマッチしたものは，処理結果として明示的に利用者に提示されるため注意が向きやすいが，マッチしなかったものには注意が行きにくいので，特に注意が必要である．

4.4 データを批判的に検討する

一般に，コーパスから肯定証拠は得られても否定証拠は得られない．ある表現が検索されなかったからと言って，それが使用できない（あるいは，使用されたことがない）と判断することはできない．肯定的な証拠を中心としたデータにより仮説が検証できる場合もあるが，否定証拠が必要とされるケースは多い．コーパスからは否定証拠は得られないため，そのような場合，母語話者による内省を利用せざるをえない．

もちろん，コーパスから直接的な否定証拠が得られなくても，間接的な否定証拠は得られることはある．しかし，どの程度の規模のコーパスであれば，間接的否定証拠が有効なのか，研究者間で共有されている基準が存在するわけではなく，どのような場合に間接的否定証拠が利用できるのかは定かではない．同様のことが肯定証拠についても言える．具体例を基に考えることにしよう．

地域的変種の差や通時的な変化を見るために，複数のコーパスにおける頻度を比較することがある．例えば，Brown，LOB，Frown，FLOB（約 100 万語）の 4 つのコーパスで *estimated* を含む ABTW 構文の数を調べると次のようになる．

(33)

Brown (AmE, 1961)	LOB (BrE, 1961)
Frown (AmE, 1992)	FLOB (BrE, 1991)

7	0
9	10

この数値だけを見ていると，1960 年代のアメリカ英語（Brown，7 例）とイギリス英語（LOB，0 例）の間，また，同じイギリス英語でも 1960 年代（LOB，0 例）と 1990 年代（FLOB，10 例）の間に，意味のある違いがあるように思えてくるが，しかし，(34) に示したように，他の語を含む例の数も見てみると，これらの差は単なる偶然（あるいは，アメリカ英語とイギリス英語の違い，時代の違いとは別の要因）によるものに見えてくる．事実，他の資料を見ると，1961 年頃のイギリス英語でも *estimated* を含む表現が見つかり，(33) の数値から ABTW 構文に関して，何か結論を導き出すのは危険であることがわかる．

(34)

		estimated		further		additional		合計	
Brown	LOB	7	0	1	11	6	4	14	15
Frown	FLOB	9	10	0	23	7	0	16	33

上記の例についてはコーパスの規模に問題があり，大規模コーパスに基づく数値であれば，このような問題は生じないかというと，そうでもない．(35) に示したのは N *after* N 構文の変種だが，(35a) は異なる名詞が使われているもの，(35b) は名詞ではなく形容詞が *after* で連結されているものである．

(35) a. How big is the Burcham family? I met auntie after cousin after brother after sister! [Web]
b. Surprising after surprising events delight the audience and a pattern starts to emerge. [Web]

どちらのタイプの変種も，The British National Corpus（約 1 億語）では該当例は見つからない．The Bank of English（5 億語以上）では，a. のタイプのみ 4 例見つかる．検索方法の問題もあるので，存在しない，あるいは 4 例のみと断言することはできないが，あっても数は極めて少ないと思われる．サーチエンジン Google による検索では，(35) の例のようにある程度の数の該当例が見つかるが，母語話者 5 人（大学教員；アメリカ人 2，オーストラリア人 2，イギリス人 1）に確認したところ，個人や個々の用例によって異なるところはあるが，概ね，(35a) タイプは容認可能だが，(35b) タイプは容認しないか容認度は低くなるという判断であった．このように，大規模コーパスにおける出現頻度数であっても，出現頻度数のみから機械的に適格性が判断できるわけではない．

少なくとも今のところ，i) どういう言語現象（構文），変種であれば，ii）どういうジャンル，規模のコーパスで，iii) どのような数値（e.g. 出現頻度，MI-score）が得られれば，仮説が支持／棄却されると言ってよいかの基準はないため，仮説に合ったデータが得られれば証拠として使い，合わないデータが得られた場合は，コーパスのジャンル，規模などが適切でないとの理由で採用しないというように，恣意的に利用される危険性もある．コーパスデータが信頼できるように見えるのは，実は，研究者が，理論の目でデータを解釈したり，母語話者としての言語直感，あるいは研究者としての言語感覚によって，データを取捨選択して利用しているからという可能性もある．「コーパスデータの客観性」を強調するあまり，このような側面が隠させてしまうと，問題の検討自体が行われなくなる危険性がある．

上で見たように，コーパスを利用するメリットは大きいのだから，問題がある（かもしれない）からと言って，利用すべきでないということには勿論ならないが，コーパスから得られたデータがどのようなものであるのか，常に批判的に検討することは重要なことであろう．

5. おわりに

直接観察不可能な文法の中身を探るには，文法を用いた活動の結果生じる観察可能なデータを

利用するしかないが，活動の種類により反映されやすい文法の側面も異なり，また，文法外のどのような要因に影響されやすいかも異なる．コーパスにせよ，内省による文法性の判断にせよ，言語知識そのものではなく，言語知識を用いた活動の結果（を集積したもの）であるのだから，そのことをよく理解した上で利用する必要がある．

コーパスをうまく活用すれば，内省などからは得にくい情報が得られるが，そのためには，分析的視点に加え，コーパスの内容の把握，情報処理の技術の習得が不可欠である．また，コンピュータによる処理では，直接データを見ていれば気付くようなことにも気付かない可能性が高くなるため，直接データを観察し，適切な言語感覚を身に付け，想像力を養うことが重要になってくる．最近では，“ユーザーフレンドリー”なツールの普及により，コーパスやツールに関してしっかりとした知識がなくともコーパスが利用できる状況になってきているが，このような状況には弊害も多く，文法研究とコーパスの関係，コーパスとツール，コーパスから得られるデータの性質をよく理解した上でコーパスを利用する必要がある．

註

1）MI-score は特定の 2 語間について連想関係の強さを計る尺度で，t-score は特定の 2 語間に何らかの連想関係があることを主張することができる確信度を計る尺度（cf. 齊藤他 2005：213，215）．具体的な計算式については，同書 pp. 132–133 を参照．

2）The Bank of English，約 5 億 2,400 万語（2008 年 5 月 10 日現在）．The British National Corpus，約 1 億語．用例に付けられた [BoE] は前者からの例，[BNC] は後者からの例であることを示す．[Web] とあるのはウェブページから採取した例（URL は省略）．用例に付いている下線はすべて大名によるもの．

3）容認度の段階性については単純化した図となっている．

4）可能な文 i，可能な文 ii，... の間に重複がない図となっているが，排他的関係にあるという意味ではない．

5）内省により利用可能な文の範囲は「可能な文」の集合と同じではない．内省による文法性の判断は運用であり，言語能力を直接反映したものではない．本稿でも具体例を挙げて指摘しているとおり，言語直感の優れた母語話者であっても，誤った判断をすることはそう珍しいことではない．

6）(25e) は異なる名詞が 3 つ現れるケースであるが，この場合には，(25e) の i. と ii. に示した 2 種類の解釈が可能である．ii. のように解釈可能な例としては，(23b) が挙げられる．

7）ここでいう「周辺的」には「瑣末な」「重要性の低い」などの含意はない．「周辺」の重要性については，梶田（2004）を参照．

8）専門家の経験的・感覚的判断とは独立に，問題の手法の処理結果から得られる予測を検証することによって，その手法の有効性を判断する可能性は考えられるが，これまでのところ，このような方法により有効性が示された手法というものは，寡聞にして知らない．

参照文献

荒木一雄，安井稔（編）．1992.『現代英文法辞典』三省堂．

Jackendoff, R. 1977. *$\bar{X}$ Syntax: A Study of Phrase Structure*. Cambridge: MIT Press.

———. 2008. “*Construction after Construction* and its Theoretical Challenges.” *Language* 84(1)：8–28.

Jespersen, O. 1909–49. *A Modern English Grammar on Historical Principles*. 7 vols. London: George Allen & Unwin.

梶田優．2004.「〈周辺〉〈例外〉は周辺・例外か」『日本語文法』4(2)：3–23.

Ohna, T. 2003. “*A Beautiful Two Weeks*: Its Syntactic Structure and the Semantic Relations of the Adjective to the Nu-

meral and Head Noun.” In S. Chiba *et al*. (eds.) *Empirical and Theoretical Investigations into Language: A Festschrift for Masaru Kajita*, 577–587. Tokyo: Kaitakusha.
大名力．2004.「コーパスからデータが得やすい構文，得にくい構文—*a beautiful two weeks* と *book after book* を例に」『英語コーパス研究』11：185–198.
齊藤俊雄・中村純作・赤野一郎（編著）．2005.『英語コーパス言語学—基礎と実践』（改訂新版）研究社.
滝沢直宏．2009.「wildly の意味と用法」（「コーパスで学ぶ本物の英語」No. 44）．*Asahi Weekly* 1852，2009 年 1 月 25 日号：21．朝日新聞社.